

# Algorithmic statistics: normal objects and universal models

Alexey Milovanov  
Moscow State University  
almas239@gmail.com

*To my school teacher V.S. Shulman, to his 70th birthday*

## Abstract

Kolmogorov suggested to measure quality of a statistical hypothesis (a model)  $P$  for a data  $x$  by two parameters: Kolmogorov complexity  $C(P)$  of the hypothesis and the probability  $P(x)$  of  $x$  with respect to  $P$ . The first parameter measures how simple the hypothesis  $P$  is and the second one how it fits. The paper [2] discovered a small class of models that are universal in the following sense. Each hypothesis  $S_{ij}$  from that class is identified by two integer parameters  $i, j$  and for every data  $x$  and for each complexity level  $\alpha$  there is a hypothesis  $S_{ij}$  with  $j \leq i \leq l(x)$  of complexity at most  $\alpha$  that has almost the best fit among all hypotheses of complexity at most  $\alpha$ . The hypothesis  $S_{ij}$  is identified by  $i$  and the leading  $i - j$  bits of the binary representation of the number of strings of complexity at most  $i$ . On the other hand, the initial data  $x$  might be completely irrelevant to the the number of strings of complexity at most  $i$ . Thus  $S_{ij}$  seems to have some information irrelevant to the data, which undermines Kolmogorov's approach: the best hypotheses should not have irrelevant information.

To restrict the class of hypotheses for a data  $x$  to those that have only relevant information, the paper [10] introduced a notion of a strong model for  $x$ : those are models for  $x$  whose total conditional complexity conditional to  $x$  is negligible. An object  $x$  is called normal if for each complexity level  $\alpha$  at least one its best fitting model of that complexity is strong.

In this paper we show that there are “many types” of normal strings (Theorem 10). Our second result states that there is a normal object  $x$  such that all its best fitting models  $S_{ij}$  are not strong for  $x$ . Our last result states that every best fit strong model for a normal object is again a normal object.

**Keywords:** algorithmic statistics, minimum description length, stochastic strings, total conditional complexity, sufficient statistic, denoising

# 1 Introduction

Let us recall the basic notion of algorithmic information theory and algorithmic statistics (see [7, 5, 9] for more details). As objects, we consider strings over the binary alphabet  $\{0, 1\}$ . The set of all strings is denoted by  $\{0, 1\}^*$  and the length of a string  $x$  is denoted by  $l(x)$ . The empty string is denoted by  $\Lambda$ .

## 1.1 Algorithmic information theory

Let  $D$  be a partial computable function mapping pairs of strings to strings. *Conditional Kolmogorov complexity* with respect to  $D$  is defined as

$$C_D(x|y) = \min\{l(p) \mid D(p, y) = x\}.$$

In this context the function  $D$  is called a *description mode* or a *decompressor*. If  $D(p, y) = x$  then  $p$  is called a *description of  $x$  conditional to  $y$*  or a *program mapping  $y$  to  $x$* .

A decompressor  $D$  is called *universal* if for every other decompressor  $D'$  there is a string  $c$  such that  $D'(p, y) = D(cp, y)$  for all  $p, y$ . By Solomonoff—Kolmogorov theorem universal decompressors exist. We pick arbitrary universal decompressor  $D$  and call  $C_D(x|y)$  the *Kolmogorov complexity* of  $x$  conditional to  $y$ , and denote it by  $C(x|y)$ . Then we define the unconditional Kolmogorov complexity  $C(x)$  of  $x$  as  $C(x|\Lambda)$ .

By  $\log n$  we denote binary logarithm. *Symmetry of information*:  $C(x) + C(y|x) \approx C(y) + C(x|y) \approx C(x, y)$ . This equality holds with accuracy  $O(\log(C(x) + C(y)))$  and is due to Kolmogorov and Levin.

## 1.2 Algorithmic statistics: basic notions

Algorithmic statistics studies explanations of observed data that are suitable in the algorithmic sense: an explanation should be simple and capture all the algorithmically discoverable regularities in the data. The data is encoded, say, by a binary string  $x$ . In this paper we consider explanations (statistical hypotheses) of the form “ $x$  was drawn at random from a finite set  $A$  with uniform distribution”.

Kolmogorov suggested in a talk [4] in 1974 to measure the quality of an explanation  $A \ni x$  by two parameters: Kolmogorov complexity  $C(A)$ <sup>1</sup> of  $A$  and the log-cardinality  $\log_2 |A|$  of  $A$ . The smaller  $C(A)$  is the simpler the explanation is. The log-cardinality measures the *fit* of  $A$ —the lower is  $|A|$  the more  $A$  fits as an explanation for any of its elements. For each complexity level  $m$  any model  $A$  for  $x$  with smallest  $\log |A|$  among models of complexity at most  $m$  for  $x$  is called a *best fit hypothesis for  $x$* . The trade off between  $C(A)$  and  $\log |A|$  is represented by the *profile* of  $x$ .

<sup>1</sup>Kolmogorov complexity of  $A$  is defined as follows. We fix any computable bijection  $A \rightarrow [A]$  from the family of all finite sets to the set of binary strings, called *encoding*. Then we define  $C(A)$  as the complexity  $C([A])$  of the code  $[A]$  of  $A$ . In a similar we define Kolmogorov complexity of other finite objects.

*Definition 1.* The *profile* of a string  $x$  is the set  $P_x$  consisting of all pairs  $(m, l)$  of natural numbers such that there exists a finite set  $A \ni x$  with  $C(A) \leq m$  and  $\log_2 |A| \leq l$ .

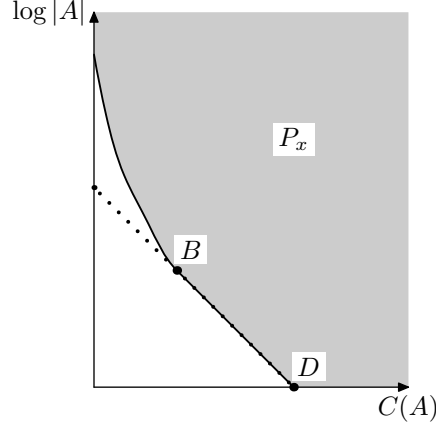


Figure 1: The profile  $P_x$  of a string  $x$ .

Both parameters  $C(A)$  and  $\log |A|$  cannot be very small simultaneously unless the string  $x$  has very small Kolmogorov complexity. Indeed,  $C(A) + \log_2 |A| \gtrsim C(x)$ , since  $x$  can be specified by  $A$  and its index in  $A$ . A model  $A \ni x$  is called *sufficient* or *optimal* if  $C(A) + \log |A| \approx C(x)$ . The value

$$\delta(x|A) = C(A) + \log |A| - C(x)$$

is called the *optimality deficiency* of  $A$  as a model for  $x$ . On Fig. 1 parameters of sufficient statistics lie on the segment  $BD$ . A sufficient statistic that has the minimal complexity is called *minimal* (MSS), its parameters are represented by the point  $B$  on Fig. 1.

*Example 2.* Consider a string  $x \in \{0, 1\}^{2^n}$  such that leading  $n$  bits of  $x$  are zeros, and the remaining bits are random, i. e.  $C(x) \approx n$ . Consider the model  $A$  for  $x$  that consists of all strings from  $\{0, 1\}^{2^n}$  that have  $n$  leading zeros. Then  $C(A) + \log |A| = \log n + O(1) + n \approx C(x)$ , hence  $A$  is a sufficient statistic for  $x$ . As the complexity of  $A$  is negligible,  $A$  is a minimal sufficient statistic for  $x$ .

The string from this example has a sufficient statistic of negligible complexity. Such strings are called *stochastic*. Are there strings that have no sufficient statistics of negligible complexity? The positive answer to this question was obtained in [8]. Such strings are called *non-stochastic*. Moreover, under some natural constraints for every set  $P$  there is a string whose profile is close to  $P$ . The constraints are listed in the following theorem:

**Theorem 3.** *Let  $x$  be a string of length  $n$  and complexity  $k$ . Then  $P_x$  has the following properties:*

- 1)  $(k + O(\log n), 0) \in P_x$ .
- 2)  $(O(\log n), n) \in P_x$ .
- 3) if  $(a, b + c) \in P_x$  then  $(a + b + O(\log n), c) \in P_x$ .
- 4) if  $(a, b) \in P_x$  then  $a + b > k - O(\log n)$ .

In other words, with logarithmic accuracy, the boundary of  $P_x$  contains a point  $(0, a)$  with  $a \leq l(x)$ , contains the point  $(C(x), 0)$ , decreases with the slope at least  $-1$  and lies above the line  $C(A) + \log |A| = C(x)$ . Conversely, given a curve with these property that has low complexity one can find a string  $x$  of length  $n$  and complexity about  $k$  such that the boundary of  $P_x$  is close to that curve:

**Theorem 4** ([11]). *Assume that we are given  $k, n$  and an upward closed set  $P$  of pairs of natural numbers such that  $(0, n), (k, 0) \in P$ ,  $(a, b + c) \in P \Rightarrow (a + c, b) \in P$  and  $(a, b) \in P \Rightarrow a + b \geq k$ . Then there is a string  $x$  of length  $n$  and complexity  $k + O(\log n)$  whose profile is  $C(P) + O(\log n)$ -close to  $P$ . (We call subsets of  $\mathbb{N}^2$   $\varepsilon$ -close if each of them is in the  $\varepsilon$ -neighborhood of the other.) By  $C(P)$  we denote the Kolmogorov complexity of the boundary of  $P$ , which is a finite object.*

### 1.3 Models of restricted type

It turns out that Theorems 3 and 4 remain valid (with smaller accuracy) even if we restrict the class of models under consideration to models from a class  $\mathcal{A}$  provided the class  $\mathcal{A}$  has the following properties.

- (1) The family  $\mathcal{A}$  is enumerable. This means that there exists an algorithm that prints elements of  $\mathcal{A}$  as lists of strings, with some separators (saying where one element of  $\mathcal{A}$  ends and another one begins).
- (2) For every  $n$  the class  $\mathcal{A}$  contains the set  $\{0, 1\}^n$ .
- (3) There exists some polynomial  $p$  with the following property: for every  $A \in \mathcal{A}$ , for every natural  $n$  and for every natural  $c < |A|$  the set of all  $n$ -bit strings in  $A$  can be covered by at most  $p(n) \cdot |A|/c$  sets of cardinality at most  $c$  from  $\mathcal{A}$ .

Any family of finite sets of strings that satisfies these three conditions is called *acceptable*.

Let us define the *profile of  $x$  with respect to  $\mathcal{A}$* :

$$P_x^{\mathcal{A}} = \{(a, b) \mid \exists A \ni x : A \in \mathcal{A}, C(A) \leq a, \log |A| \leq b\}.$$

Obviously  $P_x^{\mathcal{A}} \subseteq P_x$ . Let us fix any acceptable class  $\mathcal{A}$  of models.

**Theorem 5** ([12]). *Let  $x$  be a string of length  $n$  and complexity  $k$ . Then  $P_x^{\mathcal{A}}$  has the following properties:*

- 1)  $(k + O(\log n), 0) \in P_x^{\mathcal{A}}$ .
- 2)  $(O(\log n), n) \in P_x^{\mathcal{A}}$ .
- 3) if  $(a, b + c) \in P_x^{\mathcal{A}}$  then  $(a + b + O(\log n), c) \in P_x^{\mathcal{A}}$ .
- 4) if  $(a, b) \in P_x^{\mathcal{A}}$  then  $a + b > k - O(\log n)$ .

**Theorem 6** ([12]). Assume that we are given  $k, n$  and an upward closed set  $P$  of pairs of natural numbers such that  $(0, n), (k, 0) \in P$ ,  $(a, b+c) \in P \Rightarrow (a+c, b) \in P$  and  $(a, b) \in P \Rightarrow a+b \geq k$ . Then there is a string  $x$  of length  $n$  and complexity  $k + O(\log n)$  such that both sets  $P_x^A$  and  $P_x$  are  $C(P) + O(\sqrt{n \log n})$ -close to  $P$ .

*Remark 7.* Originally, the conclusion of Theorem 6 stated only that the set  $P_x^A$  is close to the given set  $P$ . However, as observed in [9], the proof from [12] shows also that  $P_x$  is close to  $P$ .

## 1.4 Universal models

Assume that  $A$  is sufficient statistic  $A$  for  $x$ . Then  $A$  provides a two-part code  $y = (\text{the shortest description of } A, \text{the index of } x \text{ in } A)$  for  $x$  whose total length is close to the complexity of  $x$ . The symmetry of information implies that  $C(y|x) \approx C(y) + C(x|y) - C(x)$ . Obviously, the term  $C(x|y)$  here is negligible and  $C(y)$  is at most its total length, which by assumption is close  $C(x)$ . Thus  $C(y|x) \approx 0$ , that is,  $x$  and  $y$  have almost the same information. That is, the two-part code  $y$  for  $x$  splits the information from  $x$  in two parts: the shortest description of  $A$ , the index of  $x$  in  $A$ . The second part of this two-part code is incompressible (random) conditional to the first part (as otherwise, the complexity of  $x$  would be smaller than the total length of  $y$ ). Thus the second part of this two-part code can be considered as accidental information (noise) in the data  $x$ . In a sense every sufficient statistic  $A$  identifies about  $C(x) - C(A)$  bits of accidental information in  $x$ . And thus any minimal sufficient statistic for  $x$  extracts almost all useful information from  $x$ .

However, it turns out that this viewpoint is inconsistent with the existence of universal models, discovered in [2]. Let  $L_m$  denote the list of strings of complexity at most  $m$ . Let  $p$  be an algorithm that enumerates all strings of  $L_m$  in some order. Notice that there is such algorithm of complexity  $O(\log m)$ . Denote by  $\Omega_m$  the cardinality of  $L_m$ . Consider its binary representation, i. e., the sum:

$$\Omega_m = 2^{s_1} + 2^{s_2} + \dots + 2^{s_t}, \text{ where } s_1 > s_2 > \dots > s_t.$$

According to this decomposition and  $p$ , we split  $L_m$  into groups: first  $2^{s_1}$  elements, next  $2^{s_2}$  elements, etc. Let us denote by  $S_{m,s}^p$  the group of size  $2^s$  from the partition. Notice that  $S_{m,s}^p$  is defined only for  $s$  that correspond to ones in the binary representation of  $\Omega_m$ , so  $m \geq s$ .

If  $x$  is a string of complexity at most  $m$ , it belongs to some group  $S_{m,s}^p$  and this group can be considered as a model for  $x$ . We may consider different values of  $m$  (starting from  $C(x)$ ). In this way we get different models  $S_{m,s}^p$  for the same  $x$ . The complexity of  $S_{m,s}^p$  is  $m - s + O(\log m + C(p))$ . Indeed, chop  $L_m$  into portions of size  $2^s$  each, then  $S_{m,s}^p$  is the last full portion and can be identified by  $m, s$  and the number of full portions, which is less than  $\Omega_m/2^s < 2^{m-s+1}$ . Thus if  $m$  is close to  $C(x)$  and  $C(p)$  is small then  $S_{m,s}^p$  is a sufficient statistic for  $x$ . More specifically  $C(S_{m,s}^p) + \log |S_{m,s}^p| = C(S_{m,s}^p) + s = m + O(\log m + C(p))$ .

For every  $m$  there is an algorithm  $p$  of complexity  $O(\log m)$  that enumerates all strings of complexity at most  $m$ . We will fix for every  $m$  any such algorithm  $p_m$  and denote  $S_{m,s}^{p_m}$  by  $S_{m,s}$ .

The models  $S_{m,s}$  were introduced in [2]. The models  $S_{m,s}^p$  are universal in the following sense:

**Theorem 8** ([11]). <sup>2</sup> *Let  $A$  be any finite set of strings containing a string  $x$  of length  $n$ . Then for every  $p$  there are  $s \leq m \leq n + O(1)$  such that*

- 1)  $x \in S_{m,s}^p$ ,
- 2)  $C(S_{m,s}^p | A) = O(\log n + C(p))$  (and hence  $C(S_{m,s}^p) \leq C(A) + O(\log n + C(p))$ ),
- 3)  $\delta(x | S_{m,s}^p) \leq \delta(x | A) + O(\log n + C(p))$ .

It turns out that the model  $S_{m,s}^p$  has the same information as the number  $\Omega_{m-s}$ :

**Lemma 9** ([11]). *For every  $a \leq b$  and for every  $s \leq m$ :*

- 1)  $C(\Omega_a | \Omega_b) = O(\log b)$ .
- 2)  $C(\Omega_{m-s} | S_{m,s}^p) = O(\log m + C(p))$  and  $C(S_{m,s}^p | \Omega_{m-s}) = O(\log m + C(p))$ .
- 3)  $C(\Omega_a) = a + O(\log a)$ .

By Theorem 8 for every data  $x$  there is a minimal sufficient statistic for  $x$  of the form  $S_{m,s}$ . Indeed, let  $A$  be any minimal sufficient statistic for  $x$  and let  $S_{m,s}$  be any model for  $x$  that exists by Theorem 8 for this  $A$ . Then by item 3 the statistic  $S_{m,s}$  is sufficient as well and by item 2 its complexity is also close to minimum. Moreover, since  $C(S_{m,s} | A)$  is negligible and  $C(S_{m,s}) \approx C(A)$ , by symmetry of information  $C(A | S_{m,s})$  is negligible as well. Thus  $A$  has the same information as  $S_{m,s}$ , which has the same information as  $\Omega_{m-s}$  (Lemma 9(2)). Thus if we agree that every minimal sufficient statistic extracts all useful information from the data, we must agree also that that information is the same as the information in the number of strings of complexity at most  $i$  for some  $i$ .

## 1.5 Total conditional complexity and strong models

The paper [10] suggests the following explanation to this paradox. Although conditional complexities  $C(S_{m,s} | A)$  and  $C(S_{m,s} | x)$  are small, the short programs that map  $A$  and  $x$ , respectively, to  $S_{m,s}$  work in a huge time. A priori their work time is not bounded by any total computable function of their input. Thus it may happen that practically we are not able to find  $S_{m,s}$  (and also  $\Omega_{m-s}$ ) from a MSS  $A$  for  $x$  or from  $x$  itself.

Let us consider now programs whose work time is bounded by a total computable function for the input. We get the notion of *total conditional complexity*  $CT(y | x)$ , which is the length of the shortest *total* program that maps  $x$  to  $y$ . Total conditional complexity can be much greater than plain one, see for example [6]. Intuitively, good sufficient statistics  $A$  for  $x$  must have not only

<sup>2</sup>This theorem was proved in [11, Theorem VIII.4] with accuracy  $O(\max\{\log C(y) \mid y \in A\} + C(p))$  instead of  $O(\log n)$ . Applying [11, Theorem VIII.4] to  $A' = \{y \in A \mid l(y) = n\}$  we obtain the theorem in the present form.

negligible conditional complexity  $C(A|x)$  (which follows from definition of a sufficient statistic) but also negligible *total* conditional complexity  $CT(A|x)$ . The paper [10] calls such models *A strong models for x*.

Is it true that for some  $x$  there is no **strong** MSS  $S_{m,s}$  for  $x$ ? The positive answer to this question was obtained in [10]: there are strings  $x$  whose all minimal sufficient statistics are not strong for  $x$ . Such strings are called *strange*. In particular, if  $S_{m,s}$  is a MSS for strange string  $x$  then  $CT(S_{m,s}|x)$  is large. However, a strange string has no strong MSS at all. An interesting question is whether there are strings  $x$  that do have strong MSS but have no strong MSS of the form  $S_{m,s}$ ? This question was left open in [10]. In this paper we answer this question in positive. Moreover, we show that there is a “normal” string  $x$  that has no strong MSS of the form  $S_{m,s}$  (Theorem 13). A string  $x$  is called *normal* if for every complexity level  $i$  there is a best fitting model  $A$  for  $x$  of complexity at most  $i$  (whose parameters thus lie on the border of the set  $P_x$ ) that is strong. In particular, every normal string has a strong MSS.

Our second result answers yet another question asked in [10]. Assume that  $A$  is a strong MSS for a normal string  $x$ . Is it true that the code  $[A]$  of  $A$  is a normal string itself? Our Theorem 17 states that this is indeed the case. Notice that by a result of [10] the profile  $P_{[A]}$  of  $[A]$  can be obtained from  $x$  by putting the origin in the point corresponding to parameters of  $A$  (i.e. in the point  $B$  on Fig. 1).

Our last result (which comes first in the following exposition) states that there are normal strings with any given profile, under the same restrictions as in Theorem 3 (Theorem 10 in Section 2).

## 2 Normal strings with a given profile

In this section we prove an analogue of Theorem 4 for normal strings. We start with a rigorous definitions of strong models and normal strings.

A set  $A \ni x$  is called  *$\varepsilon$ -strong statistic (model) for a string  $x$*  if  $CT(A|x) < \varepsilon$ . To represent the trade off between size and complexity of  $\varepsilon$ -strong models for  $x$  consider the  *$\varepsilon$ -strong profile of  $x$* :

$$P_x^\varepsilon = \{(a, b) \mid \exists A \ni x : CT(A|x) \leq \varepsilon, C(A) \leq a, \log |A| \leq b\}.$$

It is not hard to see that the set  $P_x^\varepsilon$  satisfies the item (3) from Theorem 3:

$$\text{for all } x \in \{0, 1\}^n \text{ if } (a, b + c) \in P_x^\varepsilon \text{ then } (a + b + O(\log n), c) \in P_x^{\varepsilon + O(\log n)}.$$

It follows from the definition that  $P_x^\varepsilon \subset P_x$  for all  $x, \varepsilon$ . Informally a string is called normal if for a negligible  $\varepsilon$  we have  $P_x \approx P_x^\varepsilon$ . Formally, for integer parameters  $\varepsilon, \delta$  we say that a string  $x$  is  *$\varepsilon, \delta$ -normal* if  $(a, b) \in P_x$  implies  $(a + \delta, b + \delta) \in P_x^\varepsilon$  for all  $a, b$ . The smaller  $\varepsilon, \delta$  are the stronger is the property of  $\varepsilon, \delta$ -normality. The main result of this section shows that for some  $\varepsilon, \delta = o(n)$  for every set  $P$  satisfying the assumptions of Theorem 3 there is an  $\varepsilon, \delta$ -normal string of length  $n$  with  $P_x \approx P$ :

**Theorem 10.** *Assume that we are given an upward closed set  $P$  of pairs of natural numbers satisfying assumptions of Theorem 4. Then there is an  $O(\log n), O(\sqrt{n \log n})$ -normal string  $x$  of length  $n$  and complexity  $k + O(\log n)$  whose profile  $P_x$  is  $C(P) + O(\sqrt{n \log n})$ -close to  $P$ .*

*Proof.* We will derive this theorem from Theorem 6. To this end consider the following family  $\mathcal{B}$  of sets. A set  $B$  is in this family if it has the form

$$B = \{uv \mid v \in \{0, 1\}^m\},$$

where  $u$  is an arbitrary binary string and  $m$  is an arbitrary natural number. Obviously, the family  $\mathcal{B}$  is acceptable, that is, it satisfies the properties (1)–(3) from Section 1.3.

Note that for every  $x$  and for every  $A \ni x$  from  $\mathcal{B}$  the total complexity of  $A$  given  $x$  is  $O(\log n)$ . So  $P_x^{\mathcal{B}} \subseteq P_x^{O(\log n)}$ . By Theorem 6 there is a string  $x$  such that  $P_x$  and  $P_x^{\mathcal{B}}$  are  $C(P) + O(\sqrt{n \log n})$ -close to  $P$ . Since  $P_x^{\mathcal{B}} \subseteq P_x^{O(\log n)} \subseteq P_x$  we conclude that  $x$  is  $O(\log n), O(\sqrt{n \log n})$ -normal.  $\square$

The proof of Theorem 10 is based on a technically difficult Theorem 6. However, for some sets  $P$  it can be shown directly with a better accuracy of  $O(\log n)$  in place of  $O(\sqrt{n \log n})$ . For instance, this happens for the smallest set  $P$ , satisfying the assumptions of Theorem 6, namely for the set

$$P = \{(m, l) \mid m \geq k, \text{ or } m + l \geq n\}.$$

Strings with such profile are called “antistochastic”.

*Definition 11.* A string  $x$  of length  $n$  and complexity  $k$  is called  $\varepsilon$ -antistochastic if for all  $(m, l) \in P_x$  either  $m > k - \varepsilon$ , or  $m + l > n - \varepsilon$ .

We will need later the fact that for every  $n$  there is an  $O(\log n)$ -antistochastic string  $x$  of length  $n$  and that such strings are normal:

**Lemma 12** (Proved in Appendix). *For all  $n$  and all  $k \leq n$  there is an  $O(\log n)$ -antistochastic string  $x$  of length  $n$  and complexity  $k + O(\log n)$ . Any such string  $x$  is  $O(\log n), O(\log n)$ -normal.*

### 3 Normal strings without universal MSS

Our main result of this section is Theorem 13 which states that there is a normal string  $x$  such that no set  $S_{m,l}$  is not a strong MSS for  $x$ .

**Theorem 13.** *For all large enough  $k$  there exist an  $O(\log k)$ -normal string  $x$  of complexity  $3k + O(\log k)$  and length  $4k$  such that:*

- 1) *The profile  $P_x$  of  $x$  is  $O(\log k)$ -close to the gray set on Fig. 3.*
- 2) *The string  $x$  has a strong MSS. More specifically, there is an  $O(\log k)$ -strong model  $A$  for  $x$  with complexity  $k + O(\log k)$  and log-cardinality  $2k$ .*
- 3) *For all simple  $q$  and all  $m, l$  the set  $S_{m,l}^q$  cannot be a strong sufficient statistic for  $x$ . More specifically, for every  $\varepsilon$ -strong  $\varepsilon$ -sufficient model  $S_{m,l}^q$  for  $x$  of complexity at most  $k + \delta$  we have  $O(\varepsilon + \delta + C(q)) \geq k - O(\log k)$ .*



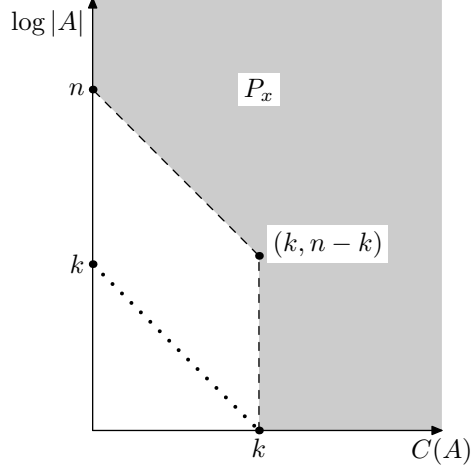


Figure 2: The profile of an  $\varepsilon$ -antistochastic string  $x$  for a small  $\varepsilon$ .

In the proof of this theorem we will need a rigorous definition of MSS and a related result from [10].

*Definition 14.* A set  $A$  is called a  $\delta, \varepsilon, D$ -minimal sufficient statistic (MSS) for  $x$  if  $A$  is an  $\varepsilon$ -sufficient statistic for  $x$  and there is no model  $B$  for  $x$  with  $C(B) < C(A) - \delta$  and  $C(B) + \log |B| - C(x) < \varepsilon + D \log C(x)$ .

The next theorem states that for every strong MSS  $B$  and for every sufficient statistic  $A$  for  $x$  the total conditional complexity  $CT(B|A)$  is negligible.

**Theorem 15** ([10], Theorem 13). *For some constant  $D$  if  $B$  is  $\varepsilon$ -strong  $\delta, \varepsilon, D$ -minimal sufficient statistic for  $x$  and  $A$  is an  $\varepsilon$ -sufficient statistic for  $x$  then  $CT(B|A) = O(\varepsilon + \delta + \log C(x))$ .*

Let us fix a constant  $D$  satisfying Theorem 15 and call a model  $\delta, \varepsilon$ -MSS if it is  $\delta, \varepsilon, D$ -MSS. Such models have the following property.

**Theorem 16** ([10], Theorem 14). *Let  $x$  be a string of length  $n$  and  $A$  be an  $\varepsilon$ -strong  $\varepsilon$ -sufficient statistic for  $x$ . Then for all  $b \geq \log |A|$  we have*

$$(a, b) \in P_x \Leftrightarrow (a + O(\varepsilon + \log n), b - \log |A| + O(\varepsilon + \log n)) \in P_{[A]}$$

*and for  $b \leq \log |A|$  we have  $(a, b) \in P_x \Leftrightarrow a + b \geq C(x) - O(\log n)$ .*

*The proof of Theorem 13.* Define  $x$  as the concatenation of strings  $y$  and  $z$ , where  $y$  is an  $O(\log k)$ -antistochastic string of complexity  $k$  and length  $2k$  (existing by Lemma 12) and  $z$  is a string of length  $2k$  such that  $C(z|y) \geq 2k - O(\log k)$  (and hence  $C(x) = 3k + O(\log k)$ ). Consider the following set  $A = \{yz' \mid l(z') = 2k\}$ . From the shape of  $P_x$  it is clear that  $A$  is an  $O(\log k), O(\log k)$ -MSS for  $x$ . Also it is clear that  $A$  is an  $O(\log k)$ -strong model for  $x$ . So, by Theorem 16

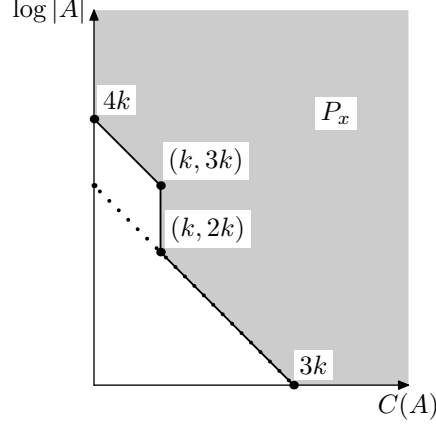


Figure 3: The profile  $P_x$  of a string  $x$  from Theorem 13.

the profile of  $x$  is  $O(\log k)$ -close to the gray set on Fig. 3. From normality of  $y$  (Lemma 12) it is not difficult to see that  $x$  is  $O(\log k)$ -normal.

Let  $S_{m,l}^q$  be an  $\varepsilon$ -strong  $\varepsilon$ -sufficient model for  $x$  of complexity at most  $k + \delta$ . We claim that  $S_{m,l}^q$  is an  $\varepsilon, \delta + O(\log k)$ -MSS for  $x$ . In other words,  $C(B) \leq C(S_{m,l}^q) - \delta - O(\log k)$  implies  $C(B) + \log |B| > C(x) + \varepsilon + D \log k$  where  $D$  is the constant from Theorem 15.

The assumption  $C(B) \leq C(S_{m,l}^q) - \delta - O(\log k)$  and the assumed upper bound for  $C(S_{m,l}^q)$  imply that  $C(B) \leq k - O(\log k)$ . From the shape of  $P_x$  it follows that  $C(B) + \log |B| \geq C(x) + k - O(\log k)$ . Notice that if  $\varepsilon$  is close to  $k$  the conclusion of the theorem is straightforward. Otherwise, the last inequality implies  $C(B) + \log |B| > C(x) + \varepsilon + D \log k$ .

By Theorem 15 we get  $CT(S_{m,l}^q | A) = O(\varepsilon + \delta + \log k)$  and thus  $CT(s_0 | y) = O(\varepsilon + \delta + \log k)$ , where  $s_0$  is the lexicographic least element in  $S_{m,l}^q$ . Denote by  $p$  a total program of length  $O(\varepsilon + \delta + \log k)$  that transforms  $y$  to  $s_0$ . Consider the following set  $B := \{p(y') \mid l(y') = 2k\}$ . We claim that if  $\varepsilon$  and  $\delta$  are not very big, then the complexity of any element from  $B$  is not greater than  $m$ . Indeed, if  $\varepsilon + \delta \leq dk$  for a small constant  $d$ , then  $l(p) < k - O(\log k)$  and hence every element from  $B$  has complexity at most  $C(B) + \log |B| + O(\log k) \leq 3k - O(\log k) \leq m$ . The last inequality holds because  $S_{m,l}^q$  is a model for  $x$  and hence  $m \geq C(x) = 3k + O(\log k)$ .

Let us run the program  $q$  until it prints all elements from  $B$ . Since  $s_0 \in B$ , there are at most  $2^l$  elements of complexity  $m$  that we have not been printed yet. So, we can find the list of all strings of complexity at most  $m$  from  $B$ ,  $q$  and some extra  $l$  bits. Since this list has complexity at least  $m - O(\log m)$  (as from this list and  $m$  we can compute a string of complexity more than  $m$ ), we get  $O(C(B) + C(q)) + l \geq m - O(\log m)$ .

Recall that the  $C(S_{m,l}^q) + \log |S_{m,l}^q|$  is equal to  $m + O(\log m + C(q))$  and is

at most  $C(x) + \varepsilon$ . Hence  $m \leq 4k$  unless  $\varepsilon > k + O(\log k + C(q))$ . Therefore the term  $O(\log m)$  in the last inequality can be re-written as  $O(\log k)$ .

Recall that the complexity of  $S_{m,l}^q$  is  $m - l + O(\log m + C(q))$ . From the shape of  $P_x$  it follows that  $C(S_{m,l}^q) \geq k - O(\log k)$  or  $C(S_{m,l}^q) + \log |S_{m,l}^q| \geq C(x) + k - O(\log k)$ . In the latter case  $\varepsilon \geq k - O(\log k)$  and we are done. In the former case  $m - l \geq k - O(\log k + C(q))$  and hence  $O(C(B) + C(q)) \geq k - O(\log k + C(q))$ .  $\square$

## 4 Hereditary of normality

In this section we prove that every strong MSS for a normal string is itself normal. Recall that a string  $x$  is called  $\varepsilon, \delta$ -normal if for every model  $B$  for  $x$  there is a model  $A$  for  $x$  with  $CT(A|x) \leq \varepsilon$  and  $C(A) \leq C(B) + \delta$ ,  $\log |A| \leq \log |B| + \delta$ .

**Theorem 17.** *There is a constant  $D$  such that the following holds. Assume that  $A$  is an  $\varepsilon$ -strong  $\delta, \varepsilon, D$ -MSS for an  $\varepsilon, \varepsilon$ -normal string  $x$  of length  $n$ . Assume that  $\varepsilon \leq \sqrt{n}/2$ . Then the code  $[A]$  of  $A$  is  $O((\varepsilon + \delta + \log n) \cdot \sqrt{n})$ -normal.*

The rest of this section is the proof of this theorem. We start with the following lemma, which is a simple corollary of Theorem 8 and Lemma 9. For the sake of completeness we prove it in Appendix.

**Lemma 18.** *For all large enough  $D$  the following holds: if  $A$  is a  $\delta, \varepsilon, D$ -MSS for  $x \in \{0, 1\}^n$  then  $C(\Omega_{C(A)}|A) = O(\delta + \log n)$ .*

We fix a constant  $D$  satisfying Lemma 18 and call a model  $\delta, \varepsilon$ -MSS if it  $\delta, \varepsilon, D$ -MSS. This  $D$  is the constant satisfying Theorem 17.

A family of sets  $\mathcal{A}$  is called *partition* if for every  $A_1, A_2 \in \mathcal{A}$  we have  $A_1 \cap A_2 \neq \emptyset \Rightarrow A_1 = A_2$ . Note that for a finite partition we can define its complexity. The next lemma states that every strong statistic  $A$  can be transformed to a strong statistic  $A_1$  such that  $A_1$  belongs to some simple partition.

**Lemma 19.** *Let  $A$  be an  $\varepsilon$ -strong statistic for  $x \in \{0, 1\}^n$ . Then there is a set  $A_1$  and a partition  $\mathcal{A}$  of complexity at most  $\varepsilon + O(\log n)$  such that:*

- 1)  $A_1$  is  $\varepsilon + O(\log n)$ -strong statistic for  $x$ .
- 2)  $CT(A|A_1) < \varepsilon + O(\log n)$  and  $CT(A_1|A) < \varepsilon + O(\log n)$ .
- 3)  $|A_1| \leq |A|$ .
- 4)  $A_1 \in \mathcal{A}$ .

*Proof.* Assume that  $A$  is an  $\varepsilon$ -strong statistic for  $x$ . Then there is a total program  $p$  such that  $p(x) = A$  and  $l(p) \leq \varepsilon$ .

We will use the same construction as in Remark 1 in [10]. For every set  $B$  denote by  $B'$  the following set:  $\{x' \in B \mid p(x') = B, x' \in \{0, 1\}^n\}$ . Notice that  $CT(A'|A)$ ,  $CT(A|A')$  and  $CT(A'|x)$  are less than  $l(p) + O(\log n) = \varepsilon + O(\log n)$  and  $|A'| \leq |A|$ .

For any  $x_1, x_2 \in \{0, 1\}^n$  with  $p(x_1) \neq p(x_2)$  we have  $p(x_1)' \cap p(x_2)' = \emptyset$ . Hence  $\mathcal{A} := \{p(x)' \mid x \in \{0, 1\}^n\}$  is a partition of complexity at most  $\varepsilon + O(\log n)$ .  $\square$

By Theorem 8 and Lemma 9 for every  $A \ni x$  there is a  $B \ni x$  such that  $B$  is informational equivalent  $\Omega_{C(B)}$  and parameters of  $B$  has are not worse than those of  $A$ . We will need a similar result for normal strings and for strong models.

**Lemma 20.** *Let  $x$  be an  $\varepsilon, \alpha$ -normal string with length  $n$  such that  $\varepsilon \leq n$ ,  $\alpha < \sqrt{n}/2$ . Let  $A$  be an  $\varepsilon$ -strong statistic for  $x$ . Then there is a set  $H$  such that:*

- 1)  $H$  is an  $\varepsilon$ -strong statistic for  $x$ .
- 2)  $\delta(x|H) \leq \delta(x|A) + O((\alpha + \log n) \cdot \sqrt{n})$  and  $C(H) \leq C(A)$ .
- 3)  $C(H|\Omega_{C(H)}) = O(\sqrt{n})$ .

*Sketch of proof (the detailed proof is deferred to Appendix).* Consider the sequence  $A_1, B_1, A_2, B_2, \dots$  of statistics for  $x$  defined as follows. Let  $A_1 := A$  and let  $B_i$  be an improvement of  $A_i$  such that  $B_i$  is informational equivalent to  $\Omega_{C(B_i)}$ , which exists by Theorem 8. Let  $A_{i+1}$  be a strong statistic for  $x$  that has a similar parameters as  $B_i$ , which exists because  $x$  is normal. (See Fig. 4.)

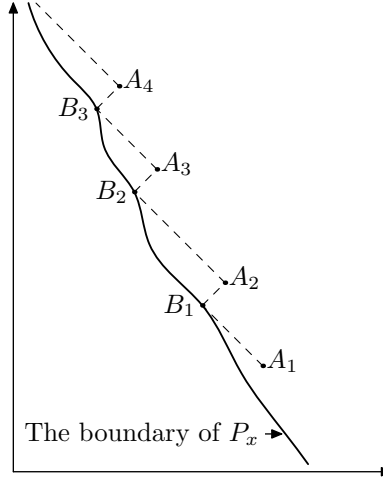


Figure 4: Parameters of statistics  $A_i$  and  $B_i$

Denote by  $N$  the minimal integer such that  $C(A_N) - C(B_N) \leq \sqrt{n}$ . For  $i < N$  the complexity of  $B_i$  is more than  $\sqrt{n}$  less than that of  $A_i$ . On the other hand, the complexity of  $A_{i+1}$  is at most  $\alpha < \sqrt{n}/2$  larger than that of  $B_i$ . Hence  $N = O(\sqrt{n})$ . Let  $H := A_N$ . By definition  $A_N$  (and  $H$ ) is strong. From  $N = O(\sqrt{n})$  it follows that the second condition is satisfied. From  $C(A_N) - C(B_N) \leq \sqrt{n}$  and definition of  $B_N$  it follows that the third condition is satisfied too (use symmetry of information).  $\square$

*Sketch of proof of Theorem 17 (the detailed proof is deferred to Appendix).* Assume that  $A$  is a  $\varepsilon$ -strong  $\delta, \varepsilon, D$ -minimal statistic for  $x$ , where  $D$  satisfies Lemma 18.

By Lemma 18  $A$  is informational equivalent to  $\Omega_{C(A)}$ . We need to prove that the profile of  $[A]$  is close to the strong profile of  $[A]$ .

Let  $\mathcal{A}$  be a simple partition and  $A_1$  a model from  $\mathcal{A}$  which exists by Lemma 19 applied to  $A, x$ . As the total conditional complexities  $CT(A_1|A)$  and  $CT(A|A_1)$  are small, the profiles of  $A$  and  $A_1$  are close to each other. This also applies to strong profiles. Therefore it suffices to show that (the code of)  $A_1$  is normal.

Let  $(a, b) \in P_{[A_1]}$ . The parameters (complexity and log-cardinality) of  $A_1$  are not larger than those of  $A$  and hence  $A_1$  is a sufficient statistic for  $x$ . By Theorem 16 we have  $(a, b + \log |A_1|) \in P_x$  (see Fig. 5).

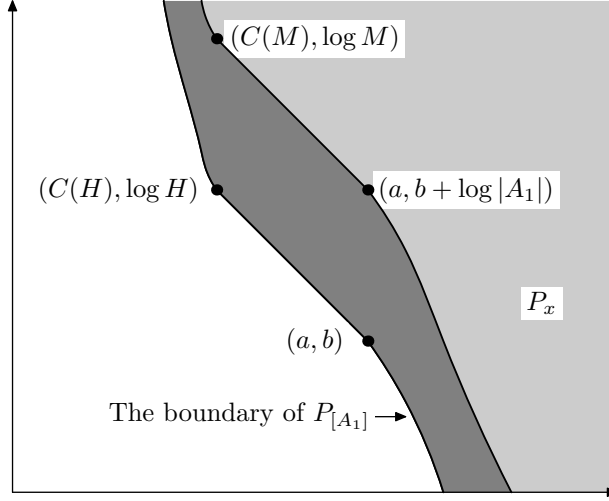


Figure 5:  $P_x$  is located  $\log |A_1|$  higher than  $P_{[A_1]}$

As  $x$  is normal, the pair  $(a, b + \log |A_1|)$  belongs to the strong profile of  $x$  as well. By Lemma 20 there is a **strong** model  $M$  for  $x$  that has low complexity conditional to  $\Omega_{C(M)}$  and whose parameters (complexity, optimality deficiency) are not worse than those of  $A_1$ .

We claim that  $C(M|A_1)$  is small. As  $A$  is informational equivalent to  $\Omega_{C(A)}$ , so is  $A_1$ . From  $\Omega_{C(A)}$  we can compute  $\Omega_{C(M)}$  (Lemma 9) and then compute  $M$  (as  $C(M|\Omega_{C(M)}) \approx 0$ ). This implies that  $C(M|A_1) \approx 0$ .

However we will need a stronger inequality  $CT(M|A_1) \approx 0$ . To find such  $M$ , we apply Lemma 19 to  $M, x$  and change it to a model  $M_1$  with the same parameters that belongs to a simple partition  $\mathcal{M}$ . Item (2) of Lemma 19 guarantees that  $M_1$  is also simple given  $A_1$  and that  $M_1$  is a strong model for  $x$ . Since  $C(M|A_1) \approx 0$ , we have  $C(M_1|A_1) \approx 0$  as well.

As  $A_1$  lies on the border line of  $P_x$  and  $C(M_1|A_1) \approx 0$ , the intersection  $A_1 \cap M_1$  cannot be much less than  $A_1$ , that is,  $\log |A_1 \cap M_1| \approx \log |A_1|$  (otherwise the model  $A_1 \cap M_1$  for  $x$  would have much smaller cardinality and almost the same complexity as  $A_1$ ). The model  $M_1$  can be computed by a total program

from  $A_1$  and its index among all  $M' \in \mathcal{M}$  with  $\log |A_1 \cap M'| \approx \log |A_1|$ . As  $\mathcal{M}$  is a partition, there are few such sets  $M'$ . Hence  $CT(M_1|A_1) \approx 0$ .

Finally, let  $H = \{A' \in \mathcal{A} \mid \log |A' \cap M_1| = \log |A_1 \cap M_1|\}$ . The model  $H$  for  $A_1$  is strong because the partition  $\mathcal{A}$  is simple and  $CT(M_1|A_1) \approx 0$ . The model  $H$  can be computed from  $M_1$ ,  $\mathcal{A}$  and  $\log |A_1 \cap M_1|$ . As  $\mathcal{A}$  is simple, we conclude that  $C(H) \lesssim C(M_1)$ . Finally  $\log |H| \leq \log |M_1| - \log |A_1|$ , because  $\mathcal{A}$  is a partition and thus it has few sets that have  $\log |A_1 \cap M_1| \approx \log |A_1|$  common elements with  $M_1$ .

Thus the complexity of  $H$  is not larger than that of  $M_1$  and the sum of complexity and cardinality of  $H$  is at most  $a + b - \log |A_1|$ . As the strong profile of  $x$  has the third property from Theorem 3, we can conclude that it includes the point  $(a, b)$ .  $\square$

## Acknowledgments

The author is grateful to N. K. Vereshchagin for advice, remarks and useful discussions.

## References

- [1] Bauwens, B., Makhlin, A., Vereshchagin, N., Zimand, M.: Short lists with short programs in short time. ECCC report TR13-007. <http://eccc.hpi-web.de/report/2013/007/>
- [2] P. Gács, J. Tromp, P.M.B. Vitányi. Algorithmic statistics, *IEEE Trans. Inform. Th.*, 47:6 (2001), 2443–2463.
- [3] Kolmogorov A. N.  
*"Three approaches to the quantitative definition of information". Problems Inform. Transmission, v. 1 (1965), no. 1, p. 1-7.*
- [4] A.N. Kolmogorov, Talk at the Information Theory Symposium in Tallinn, Estonia, 1974.
- [5] Li M., Vitányi P., *An Introduction to Kolmogorov complexity and its applications*, 3rd ed., Springer, 2008 (1 ed., 1993; 2 ed., 1997), xxiii+790 pp. ISBN 978-0-387-49820-1.
- [6] A. Shen, Game Arguments in Computability Theory and Algorithmic Information Theory. Proceedings of CiE 2012, 655–666.
- [7] A. Shen, Around Kolmogorov complexity: basic notions and results. *Measures of Complexity. Festschrift for Alexey Chervonenkis*. Editors: V. Vovk, H. Papadopoulos, A. Gammerman. Springer, 2015. ISBN: 978-3-319-21851-9

- [8] A. Shen *The concept of  $(\alpha, \beta)$ -stochasticity in the Kolmogorov sense, and its properties. Soviet Mathematics Doklady, 271(1):295–299, 1983*
- [9] A. Shen, V. Uspensky, N. Vereshchagin *Kolmogorov complexity and algorithmic randomness*. MCCME, 2013 (Russian). English translation: <http://www.lirmm.fr/~ashen/kolmbook-eng.pdf>
- [10] Nikolay Vereshchagin *"Algorithmic Minimal Sufficient Statistics: a New Approach". Theory of Computing Systems 56(2) 291-436 (2015)*
- [11] N. Vereshchagin and P. Vitányi *"Kolmogorov's Structure Functions with an Application to the Foundations of Model Selection". IEEE Transactions on Information Theory 50:12 (2004) 3265-3290. Preliminary version: Proc. 47th IEEE Symp. Found. Comput. Sci., 2002, 751–760.*
- [12] Paul Vitányi, Nikolai Vereshchagin. *"On Algorithmic Rate-Distortion Function". Proc. of 2006 IEEE International Symposium on Information Theory Sunday, July 9 -Friday, July 14, 2006 Seattle, Washington.*

## Appendix

*Proof of Lemma 12.* Let  $x$  be the lexicographic first string of length  $n$  that is not covered by any set  $A$  of cardinality  $2^{n-k}$  and complexity less than  $k$ . By a direct counting such a string exists. The string  $x$  can be computed from  $k, n$  and the number of halting programs of length less than  $k$  hence  $C(x) \leq k + O(\log n)$ . To prove that  $x$  is normal it is enough to show that for every  $i \leq k$  there is a  $O(\log n)$ -strong statistics  $A_i$  for  $x$  with  $C(A_i) \leq i + O(\log n)$  and  $\log |A_i| = n - i$ .

Let  $A_k = \{x\}$  and for  $i < k$  let  $A_i$  be the set of all strings of length  $n$  whose the first  $i$  bits are the same as those of  $x$ . By the construction  $C(A_i) \leq i + O(\log n)$  and  $\log |A_i| = n - i$ .  $\square$

*Proof of Lemma 18.* By Lemma 8 there is  $S_{k,m} \ni x$  such that:

$$C(S_{k,m}|A) = O(\log n) \text{ and } \delta(x|S_{k,m}) \leq \delta(x|A) + O(\log n). \quad (1)$$

From  $\delta(x|S_{k,m}) \leq \delta(x|A) + O(\log n)$  it follows that  $S_{k,m}$  is an  $\varepsilon + O(\log n)$ -sufficient statistic for  $x$ . If the constant  $D$  is chosen appropriately, then  $S_{k,m}$  is an  $\varepsilon + D \cdot \log n$ -sufficient statistic for  $x$ , hence, by definition of MSS:

$$C(S_{k,m}) > C(A) - \delta. \quad (2)$$

We can estimate  $C(\Omega_{C(A)}|A)$  as follows:

$$C(\Omega_{C(A)}|A) \leq C(\Omega_{C(A)}|\Omega_{C(S_{k,m})}) + C(\Omega_{C(S_{k,m})}|S_{k,m}) + C(S_{k,m}|A). \quad (3)$$

To prove the lemma it remains to show that every term of the right hand side of this inequality is  $O(\delta + \log n)$ . For the third term it follows from (1).

To prove it for the first term note that  $|C(A) - C(S_{k,m})| \leq \delta + O(\log n)$  by (1) and (2). Now the inequality  $C(\Omega_{C(A)}|\Omega_{C(S_{k,m})}) \leq \delta + O(\log n)$  follows from the following simple lemma.

**Lemma 21.** *Let  $a, b$  be some integers. Then*

$$C(\Omega_a|\Omega_b) \leq |a - b| + O(\log(a + b)).$$

*Proof.* Consider two cases. If  $b \geq a$ , then  $C(\Omega_a|\Omega_b) = O(\log b)$  by the first statement of Lemma 9.

If  $b < a$  we get  $C(\Omega_b|\Omega_a) = O(\log a)$  by the same argument. By symmetry of information we get:

$$C(\Omega_a|\Omega_b) = C(\Omega_a) - C(\Omega_b) + O(\log(a + b)).$$

To conclude the required statement it remains to recall that  $C(\Omega_a) = a + O(\log a)$  and  $C(\Omega_b) = b + O(\log b)$  by Lemma 9.  $\square$

Now it remains to show that the second term  $C(\Omega_{C(S_{k,m})}|S_{k,m})$  of the right hand side of the inequality (3) is  $O(\log n)$ . This is an easy corollary from the second item of Lemma 9 and the equality  $C(S_{k,m}) = k - m + O(\log k)$ .  $\square$

*Proof of Lemma 20.* Let  $E$  be a statistic for  $x$ . Denote by  $f(E)$  a statistic for  $x$  that is not worse than  $E$  and is equivalent to  $\Omega_t$  for some  $t$ , that is, a statistic that exists by Theorem 8:

$$\begin{aligned} C(f(E)|E) &= O(\log n), \quad \delta(x|f(E)) \leq \delta(x|E) + O(\log n), \\ C(f(E)|\Omega_{C(f(E))}) &= O(\log n), \quad C(\Omega_{C(f(E))}|f(E)) = O(\log n). \end{aligned}$$

Denote by  $g(E)$  a statistic for  $x$  such that

$$C(g(E)) < C(E) + \alpha, \quad \log |g(E)| < C(E) + \alpha, \quad g(E) \text{ is } \varepsilon\text{-strong}.$$

Such model  $g(E)$  exists for every  $E$  because  $x$  is  $\varepsilon, \alpha$ -normal.

Consider the following sequence:

$$A_1 = A, \quad B_1 = f(A_1), \quad A_2 = g(B_1), \quad B_2 = f(A_2), \dots$$

Let us call a pair  $A_i B_i$  a *big step* if  $C(A_i) - C(B_i) > \sqrt{n}$ . Denote by  $N$  the minimal integer such that  $A_N B_N$  is not a big step. Let us show that  $N = O(\sqrt{n})$ . Indeed,  $C(A_{i+1}) < C(B_i) + \alpha$  and thus  $C(A_{i+1}) - C(A_i) > \sqrt{n} - \alpha$  for every  $i < N$ . On the other hand  $C(A_1) \leq C(x) + CT(A_1|x) \leq n + \varepsilon$ . Therefore  $N \cdot (\sqrt{n} - \alpha) \leq n + \varepsilon$ . Since  $\alpha < \sqrt{n}/2$ ,  $\varepsilon \leq n$  we have  $N = O(\sqrt{n})$ .

Let  $H = A_N$ . Let us show that  $H$  satisfies all the requirements.

- 1)  $A_N$  is an  $\varepsilon$ -strong model for  $x$  by definition of  $g$ .
- 2) Let us estimate of  $\delta(x|A_N)$ . We have  $\delta(x|A_{i+1}) \leq \delta(x|B_i) + 2 \cdot \alpha$  and  $\delta(x|B_i) \leq \delta(x|A_i) + O(\log n)$  for every  $i$ . So

$$\delta(x|A_N) \leq \delta(x|A_1) + N \cdot (2\alpha + O(\log n)) \leq \delta(x|A_1) + O(\alpha + \log n) \cdot \sqrt{n}.$$

In a similar way we can estimate the complexity of  $A_N$ :  $C(B_i) < C(A_i) - \sqrt{n}$  if  $i < N$  and  $C(A_{i+1}) < C(B_i) + \alpha$ . As  $\alpha < \sqrt{n}/2$  we conclude that  $C(A_{i+1}) < C(A_i)$  for  $i < N$ . Hence  $C(A_N) \leq C(A_1)$ .



3) To estimate  $C(B_N|A_N)$  we use the following inequality:

$$C(A_N|\Omega_{C(A_N)}) \leq C(A_N|B_N) + C(B_N|\Omega_{C(B_N)}) + C(\Omega_{C(B_N)}|\Omega_{C(A_N)}).$$

It remains to show that all terms in the right hand side are equal to  $O(\sqrt{n})$ . This bound holds for the first term because, as  $A_N B_N$  is not a big step. The second term is equal to  $O(\sqrt{n})$  by the definition of  $f$ . For the third term we use the inequalities  $|C(A_N) - C(B_N)| < \sqrt{n}$  (the pair  $A_N, B_N$  is not a big step and  $C(B_N|A_N) = O(\log n)$ ) and Lemma 21.  $\square$

*Detailed proof of Theorem 17.* Again we will use the notations from the sketch of proof.

*Step 1: From  $A$  to  $A_1$ .*

As the model  $A$  is  $\varepsilon$ -strong for  $x$ , by Lemma 19 there is an  $\varepsilon + O(\log n)$ -strong statistic  $A_1$  for  $x$  that belongs to an  $\varepsilon + O(\log n)$ -simple partition  $\mathcal{A}$  such that

$$CT(A|A_1) < \varepsilon + O(\log n) \quad (4)$$

$$CT(A_1|A) < \varepsilon + O(\log n) \quad (5)$$

$$|A_1| \leq |A|. \quad (6)$$

We will show that  $P_{[A_1]}$  is close to  $P_{[A_1]}^{O((\varepsilon+\delta+\log n)\cdot\sqrt{n})}$  and then we will prove a similar statement for  $A$ .

Let  $(a, b) \in P_{[A]}$ . We need to show that  $(a, b)$  is close to  $P_{[A_1]}^{O((\varepsilon+\delta+\log n)\cdot\sqrt{n})}$ . This is straightforward, if  $a \geq C(A)$ . Therefore we will assume that  $a < C(A)$ . From (5) it is easy to see that  $(a + O(\varepsilon + \log n), b + O(\varepsilon + \log n)) \in P_{[A_1]}$ .

The set  $A$  is an  $\varepsilon$ -sufficient statistic for  $x$ . From this, (5) and (6) it follows that  $A_1$  is an  $O(\varepsilon + \log n)$ -sufficient statistic for  $x$ .

*Step 2: From  $A$  to  $M$ .*

From now on we will omit terms of the order  $O((\varepsilon + \delta + \log n) \cdot \sqrt{n})$ .

By Theorem 16 from sufficiency of  $A_1$  it follows that  $(a, b + \log |A_1|) \in P_x$ . As  $x$  is  $\varepsilon, \varepsilon$ -normal, a point with similar parameters belongs to  $P_x^\varepsilon$ . A statistic with corresponding parameters can be improved by Lemma 20, i. e., there is an  $\varepsilon$ -strong statistic  $M$  for  $x$  such that:

$$C(M|\Omega_{C(M)}) = 0, \quad C(M) \leq a, \quad (7)$$

$$\text{and } \delta(x|M) \leq a + b + \log |A_1| - C(x). \quad (8)$$

*Step 3: From  $M$  to  $M_1$ .*

By Lemma 19 we can transform  $M$  to an  $\varepsilon + O(\log n)$ -strong statistic  $M_1$  for  $x$  that belongs to an  $\varepsilon + O(\log n)$ -simple partition  $\mathcal{M}$  and whose parameters are not worse than those of  $M$ :

$$CT(M|M_1) = 0, \quad CT(M_1|M) = 0, \quad |M_1| \leq |M|, \quad (9)$$

$$C(M_1|\Omega_{C(M_1)}) = 0, \quad (10)$$

$$C(M_1) \leq a, \quad (11)$$

$$\delta(x|M_1) \leq a + b + \log |A_1| - C(x). \quad (12)$$

Now we need the following

**Lemma 22.**  $\log |A_1 \cap M_1| = \log |A_1|$  (up to  $O((\varepsilon + \delta + \log n) \cdot \sqrt{n})$ ).

*Proof of Lemma.* The model  $A$  is a  $\delta, \varepsilon, D$ -MSS for  $x$ , hence by Lemma 18  $C(\Omega_{C(A)}|A) = 0$ . On the other hand we have  $C(A|A_1) = 0$ . Hence

$$C(\Omega_{C(A)}|A_1) = 0. \quad (13)$$

Recall that we assume that  $a < C(A)$ . Inequality (11) states that  $C(M_1) < a$  and therefore  $C(M_1) < C(A)$ . Hence, from Lemma 9 it follows that

$$C(\Omega_{C(M_1)}|\Omega_{C(A)}) = 0.$$

From this, (10) and (13) it follows that  $C(M_1|A_1) = 0$ . Obviously, we have  $C(M_1 \cap A_1) \leq C(A_1) + C(M_1|A_1)$  and thus

$$C(M_1 \cap A_1) \leq C(A_1). \quad (14)$$

As  $A_1$  is a sufficient statistic for  $x$  we conclude that

$$\log |A_1 \cap M_1| + C(M_1 \cap A_1) \geq C(A_1) + \log |A_1|.$$

From this and (14) it follows that  $\log |A_1 \cap M_1| \geq \log |A_1|$ .  $\square$

*Step 4: Constructing  $H$ .*

Denote by  $H$  the family of sets from  $\mathcal{A}$  which have the same size of intersection with  $M_1$  as  $A_1$  up to a factor of 2:

$$H = \{A' \in \mathcal{A} \mid \lfloor \log |A' \cap M_1| \rfloor = \lfloor \log |A_1 \cap M_1| \rfloor\}.$$

As  $\mathcal{A}$  is partition, we have  $|H| \leq |M_1|/(2 \cdot |A_1 \cap M_1|)$ . Therefore we have

$$\log |H| \leq \log |M_1| - \log |A_1|.$$

We can compute  $H$  from  $M_1$ ,  $\mathcal{A}$  and  $\lfloor \log |A_1 \cap M_1| \rfloor$ , so:

$$C(H) \leq C(M_1) + C(\mathcal{A}) = C(M_1).$$

By a similar reason we have

$$CT(H|A_1) \leq CT(M_1|A_1) + C(\mathcal{A}) + O(1) \leq CT(M_1|A_1).$$

To estimate  $CT(M_1|A_1)$  recall that  $\mathcal{M}$  is a partition, so there are at most  $|A_1|/(2 \cdot |A_1 \cap M_1|)$  elements from  $\mathcal{M}$  who have  $|M_1 \cap A_1|/2$  common strings with  $A_1$ . Thus we have:

$$CT(H|A_1) \leq CT(M_1|A_1) \leq \log |A_1| - \log |A_1 \cap M_1| + 1 + C(\mathcal{M}) \approx 0.$$

Thus  $H$  is strong statistic for  $A_1$  of complexity at most  $C(M_1)$  and log-size at most  $\log |M_1| - \log |A_1|$ :

$$(C(M_1), \log |M_1| - \log |A_1|) \in P_{[A_1]}^{O(\sqrt{n} + \varepsilon + \delta)}. \quad (15)$$

*Step 4: Back to A.*

Inequality (11) states that  $a \geq C(M_1)$ . Since a strong profile also has the third property from Theorem 3, we can add  $a - C(M_1)$  to the first component and subtract it from the second component of the left hand side of (15) (i. e. make the statistic smaller but more complex):

$$(a, \log |M_1| - \log |A_1| - a + C(M_1)) \in P_{[A_1]}^{O(\sqrt{n} + \varepsilon + \delta)}.$$

By (12) the second component becomes less than  $b$ , i.e.  $(a, b) \in P_{[A_1]}^{O(\sqrt{n} + \varepsilon + \delta)}$ .

We have shown that there is a set  $B \ni [A_1]$  such that:

$$C(B) = a, \quad \log |B| = b, \quad CT(B|[A_1]) = O(\sqrt{n} + \varepsilon + \delta).$$

Equation (4) states that there is a total programs  $p$  of length  $\varepsilon + O(\log n)$  such that  $p([A_1]) = [A]$ . Consider the set  $D := \{p(t) | t \in B\}$ . The set  $D$  is the required model for  $A$ . Indeed, we have  $[A] \in D$ ,  $\log |D| \leq \log |B|$ ,  $C(D) \leq C(B) + l(p) + O(1) = a + O(\log n + \varepsilon)$ ,  $CT(D|[A]) \leq CT(D|B) + CT(B|A_1) + CT(A_1|A) \leq O(\sqrt{n} + \varepsilon + \delta)$ . Therefore we have

$$(a + O(\log n + \varepsilon), b + O((\varepsilon + \delta + \log n) \cdot \sqrt{n})) \in P_{[A]}^{O(\sqrt{n} + \varepsilon + \delta)}.$$

□